

Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation

Year: 2023 | Citations: 388 | Authors: Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, Danqi Chen

Abstract

The rapid progress in open-source large language models (LLMs) is significantly advancing AI development. Extensive efforts have been made before model release to align their behavior with human values, with the primary goal of ensuring their helpfulness and harmlessness. However, even carefully aligned models can be manipulated maliciously, leading to unintended behaviors, known as "jailbreaks". These jailbreaks are typically triggered by specific text inputs, often referred to as adversarial prompts. In this work, we propose the generation exploitation attack, an extremely simple approach that disrupts model alignment by only manipulating variations of decoding methods. By exploiting different generation strategies, including varying decoding hyper-parameters and sampling methods, we increase the misalignment rate from 0% to more than 95% across 11 language models including LLaMA2, Vicuna, Falcon, and MPT families, outperforming state-of-the-art attacks with \$30\times\$ lower computational cost. Finally, we propose an effective alignment method that explores diverse generation strategies, which can reasonably reduce the misalignment rate under our attack. Altogether, our study underscores a major failure in current safety evaluation and alignment procedures for open-source LLMs, strongly advocating for more comprehensive red teaming and better alignment before releasing such models. Our code is available at https://github.com/Princeton-SysML/Jailbreak_LLM.